(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
5 August 2004 (05.08.2004)

PCT

(10) International Publication Number
WO 2004/066273 A1

(51) International Patent Classification[7]: G10L 21/02, 11/02, 15/24

(21) International Application Number: PCT/EP2004/000104

(22) International Filing Date: 9 January 2004 (09.01.2004)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
03001637.2 24 January 2003 (24.01.2003) EP
03022561.9 2 October 2003 (02.10.2003) EP

(71) Applicant *(for all designated States except US)*: SONY ERICSSON MOBILE COMMUNICATIONS AB [SE/SE]; Nya Vattentornet, S-221 88 Lund (SE).

(72) Inventor; and

(75) Inventor/Applicant *(for US only)*: TANEDA, Morio [DE/DE]; Franziskanerstrasse 28, 81669 Munich (DE).

(74) Agent: RUPP, Christian; Mitscherlich & Partner, Sonnenstrasse 33, Postfach 33 06 09, 80066 München (DE).

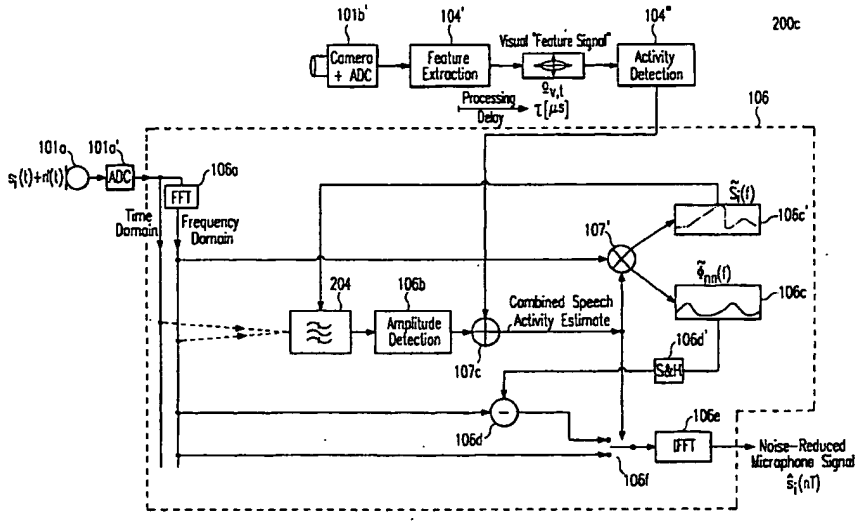
(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NA, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IT, LU, MC, NL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *with international search report*

[Continued on next page]

(54) Title: NOISE REDUCTION AND AUDIO-VISUAL SPEECH ACTIVITY DETECTION



(57) Abstract: The present invention generally relates to the field of noise reduction systems which are equipped with an audio-visual user interface, in particular to an audio-visual speech activity recognition system (200b/c) of a video-enabled telecommunication device which runs a real-time lip tracking application that can advantageously be used for a near-speaker detection algorithm in an environment where a speaker's voice is interfered by a statistically distributed background noise ($n'(t)$) including both environmental noise ($n(t)$) and surrounding persons' voices.

WO 2004/066273 A1

— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

Noise Reduction and Audio-Visual Speech Activity Detection

FIELD AND BACKGROUND OF THE INVENTION

The present invention generally relates to the field of noise reduction based on speech activity recognition, in particular to an audio-visual user interface of a telecommunication device running an application that can advantageously be used e.g. for a near-speaker detection algorithm in an environment where a speaker's voice is interfered by a statistically distributed background noise including environmental noise as well as surrounding persons' voices.

Discontinuous transmission of speech signals based on speech/pause detection represents a valid solution to improve the spectral efficiency of new-generation wireless communication systems. In this context, robust voice activity detection algorithms are required, as conventional solutions according to the state of the art present a high misclassification rate in the presence of the background noise typical of mobile environments.

A voice activity detector (VAD) aims to distinguish between a speech signal and several types of acoustic background noise even with low signal-to-noise ratios (SNRs). Therefore, in a typical telephone conversation, such a VAD, together with a comfort noise generator (CNG), is used to achieve silence compression. In the field of multimedia communications, silence compression allows a speech channel to be shared with other types of information, thus guaranteeing simultaneous voice and data applications. In cellular radio systems which are based the Discontinuous Transmission (DTX) mode, such as GSM, VADs are applied to reduce co-channel interference and power consumption of the portable equipment. Furthermore, a VAD is vital to reduce the average data bit rate in future generations of digital cellular networks such as the UMTS, which provide for a variable bit-rate (VBR) speech coding. Most of the capacity gain is due to the distinction between speech activity and in-

activity. The performance of a speech coding approach which is based on phonetic classification, however, strongly depends on the classifier, which must be robust to every type of background noise. As is well known, the performance of a VAD is critical for the overall speech quality, in particular with low SNRs. In case speech frames are detected as noise, intelligibility is seriously impaired owing to speech clipping in the conversation. If, on the other hand, the percentage of noise detected as speech is high, the potential advantages of silence compression are not obtained. In the presence of background noise it may be difficult to distinguish between speech and silence. Hence, for voice activity detection in wireless environments more efficient algorithms are needed.

Although the Fuzzy Voice Activity Detector (FVAD) proposed in „Improved VAD G.729 Annex B for Mobile Communications Using Soft Computing" (Contribution ITU-T, Study Group 16, Question 19/16, Washington, September 2-5, 1997) by F. Beritelli, S. Casale, and A. Cavallaro performs better than other solutions presented in literature, it exhibits an activity increase, above all in the presence of non-stationary noise. The functional scheme of the FVAD is based on a traditional pattern recognition approach wherein the four differential parameters used for speech activity/inactivity classification are the full-band energy difference, the low-band energy difference, the zero-crossing difference, and the spectral distortion. The matching phase is performed by a set of fuzzy rules obtained automatically by means of a new hybrid learning tool as described in „FuGeNeSys: Fuzzy Genetic Neural System for Fuzzy Modeling" by M. Russo (to appear in IEEE Transaction on Fuzzy Systems). As is well known, a fuzzy system allows a gradual, continuous transition rather than a sharp change between two values. Thus, the Fuzzy VAD returns a continuous output signal ranging from 0 (non-activity) to 1 (activity), which does not depend on whether single input signals have exceeded a predefined threshold or not, but on an overall evaluation of the values they have assumed („defuzzyfication process"). The final decision is made by comparing the output of the fuzzy system, which varies in a range between 0 and 1, with a fixed threshold experimentally chosen as described in "Voice Control of the Pan-European Digital Mobile Radio System" (ICC '89, pp. 1070-1074) by C. B. Southcott et al.

Just as voice activity detectors conventional automatic speech recognition (ASR) systems also experience difficulties when being operated in noisy environments since accuracy of

conventional ASR algorithms largely decreases in noisy environments. When a speaker is talking in a noisy environment including both ambient noise as well as surrounding persons' interfering voices, a microphone picks up not only the speaker's voice but also these background sounds. Consequently, an audio signal which encompasses the speaker's voice superimposed by said background sounds is processed. The louder the interfering sounds, the more the acoustic comprehensibility of the speaker is reduced. To overcome this problem, noise reduction circuitries are applied that take use of the different frequency regions of environmental noise and the respective speaker's voice.

A typical noise reduction circuitry for a telephony-based application based on a speech activity estimation algorithm according to the state of the art that implements a method for correlating the discrete signal spectrum $S(k\cdot\Delta f)$ of an analog-to-digital-converted audio signal $s(t)$ with an audio speech activity estimate is shown in Fig. 2a. Said audio speech activity estimate is obtained by an amplitude detection of the digital audio signal $s(nT)$. The circuit outputs a noise-reduced audio signal $\hat{s}_i(nT)$, which is calculated by subjecting the difference of the discrete signal spectrum $S(k\cdot\Delta f)$ and a sampled version $\tilde{\Phi}_{nn}(k\cdot\Delta f)$ of the estimated noise power density spectrum $\tilde{\Phi}_{nn}(f)$ of a statistically distributed background noise $\tilde{n}(t)$ to an Inverse Fast Fourier Transform (IFFT).

BRIEF DESCRIPTION OF THE STATE OF THE ART

The invention described in US 5,313,522 refers to a device for facilitating comprehension by a hearing-impaired person participating in a telephone conversation, which comprises a circuitry for converting received audio speech signals into a series of phonemes and an arrangement for coupling the circuitry to a POTS line. The circuit thereby includes an arrangement which correlates the detected series of phonemes with recorded lip movements of a speaker and displays these lip movements in subsequent images on a display device, thereby permitting the hearing-impaired person to carry out a lipreading procedure while listening to the telephone conversation, which improves the person's level of comprehension.

The invention disclosed in WO 99/52097 pertains to a communication device and a method for sensing the movements of a speaker's lips, generating an audio signal corresponding to detected lip movements of said speaker and transmitting said audio signal, thereby sensing a level of ambient noise and accordingly controlling the power level of the audio signal to be transmitted.

OBJECT OF THE UNDERLYING INVENTION

In view of the state of the art mentioned above, it is the object of the present invention to enhance the speech/pause detection accuracy of a telephony-based voice activity detection (VAD) system. In particular, it is the object of the invention to increase the signal-to-interference ratio (SIR) of a recorded speech signal in crowded environments where a speaker's voice is severely interfered by ambient noise and/or surrounding persons' voices.

The aforementioned object is achieved by means of the features in the independent claims. Advantageous features are defined in the subordinate claims.

SUMMARY OF THE INVENTION

The present invention is dedicated to a noise reduction and automatic speech activity recognition system having an audio-visual user interface, wherein said system is adapted for running an application for combining a visual feature vector $\underline{o}_{v,nT}$ that comprises features extracted from a digital video sequence $v(nT)$ showing a speaker's face by detecting and analyzing e.g. lip movements and/or facial expressions of said speaker S_i with an audio feature vector $\underline{o}_{a,nT}$ which comprises features extracted from a recorded analog audio sequence $s(t)$. Said audio sequence $s(t)$ thereby represents the voice of said speaker S_i interfered by a statistically distributed background noise

$$n'(t) = n(t) + s_{Int}(t), \qquad (1)$$

which includes both environmental noise $n(t)$ and a weighted sum of surrounding persons' interfering voices

$$s_{Int}(t) \propto \sum_{j=1}^{N} a_j \cdot s_j(t - T_j) \text{ (for } j \neq i) \quad (2a)$$

$$\text{with } a_j = \frac{1}{4\pi \cdot R_{jM}^2} \text{ [m}^{-2}\text{]} \quad (2b)$$

in the environment of said speaker S_i. Thereby, N denotes the total number of speakers (inclusive of said speaker S_i), a_j is the attenuation factor for the interference signal $s_j(t)$ of the j-th speaker S_j in the environment of the speaker S_i, T_j is the delay of $s_j(t)$, and R_{jM} denotes the distance between the j-th speaker S_j and a microphone recording the audio signal $s(t)$. By tracking the lip movement of a speaker, visual features are extracted which can then be analyzed and used for further processing. For this reason, the bimodal perceptual user interface comprises a video camera pointing to the speaker's face for recording a digital video sequence $v(nT)$ showing lip movements and/or facial expressions of said speaker S_i, audio feature extraction and analyzing means for determining acoustic-phonetic speech characteristics of the speaker's voice and pronunciation based on the recorded audio sequence $s(t)$, and visual feature extraction and analyzing means for continuously or intermittently determining the current location of the speaker's face, tracking lip movements and/or facial expressions of the speaker in subsequent images and determining acoustic-phonetic speech characteristics of the speaker's voice and pronunciation based on the detected lip movements and/or facial expressions.

According to the invention, the aforementioned extracted and analyzed visual features are fed to a noise reduction circuit that is needed to increase the signal-to-interference ratio (SIR) of the recorded audio signal $s(t)$. Said noise reduction circuit is specially adapted to perform a near-speaker detection by separating the speaker's voice from said background noise $\tilde{n}(t)$ based on the derived acoustic-phonetic speech characteristics

$$\underline{o}_{av,nT} := [\underline{o}_{a,nT}^{T}, \underline{o}_{v,nT}^{T}]^{T} \quad (3)$$

It outputs a speech activity indication signal ($\hat{s}_i(nT)$) which is obtained by a combination of speech activity estimates supplied by said audio feature extraction and analyzing means as well as said visual feature extraction and analyzing means.

BRIEF DESCRIPTION OF THE DRAWINGS

Advantageous features, aspects, and useful embodiments of the invention will become evident from the following description, the appended claims, and the accompanying drawings. Thereby,

Fig. 1 shows a noise reduction and speech activity recognition system having an audio-visual user interface, said system being specially adapted for running a real-time lip tracking application which combines visual features $o_{v,nT}$ extracted from a digital video sequence $v(nT)$ showing the face of a speaker S_i by detecting and analyzing the speaker's lip movements and/or facial expressions with audio features $o_{a,nT}$ extracted from an analog audio sequence $s(t)$ representing the voice of said speaker S_i interfered by a statistically distributed background noise $n'(t)$,

Fig. 2a is a block diagram showing a conventional noise reduction and speech activity recognition system for a telephony-based application based on an audio speech activity estimation according to the state of the art,

Fig. 2b shows an example of a camera-enhanced noise reduction and speech activity recognition system for a telephony-based application that implements an audio-visual speech activity estimation algorithm according to one embodiment of the present invention,

Fig. 2c shows an example of a camera-enhanced noise reduction and speech activity recognition system for a telephony-based application that implements an audio-visual speech activity estimation algorithm according to a further embodiment of the present invention,

Fig. 3a shows a flow chart illustrating a near-end speaker detection method reducing the noise level of a detected analog audio sequence $s(t)$ according to the embodiment depicted in Fig. 1 of the present invention,

Fig. 3b shows a flow chart illustrating a near-end speaker detection method according to the embodiment depicted in Fig. 2b of the present invention, and

Fig. 3c shows a flow chart illustrating a near-end speaker detection method according to the embodiment depicted in Fig. 2c of the present invention.

DETAILED DESCRIPTION OF THE UNDERLYING INVENTION

In the following, different embodiments of the present invention as depicted in Figs. 1, 2b, 2c, and 3a-c shall be explained in detail. The meaning of the symbols designated with reference numerals and signs in Figs. 1 to 3c can be taken from an annexed table.

According to a first embodiment of the invention as depicted in Fig. 1, said noise reduction and speech activity recognition system 100 comprises a noise reduction circuit 106 which is specially adapted to reduce the background noise $n'(t)$ received by a microphone 101a and to perform a near-speaker detection by separating the speaker's voice from said background noise $n'(t)$ as well as a multi-channel acoustic echo cancellation unit 108 being specially adapted to perform a near-end speaker detection and/or double-talk detection algorithm based on acoustic-phonetic speech characteristics derived with the aid of the aforementioned audio and visual feature extraction and analyzing means 104a+b and 106b, respectively. Thereby, said acoustic-phonetic speech characteristics are based on the opening of a speaker's mouth as an estimate of the acoustic energy of articulated vowels or diphthongs, respectively, rapid movement of the speaker's lips as a hint to labial or labio-dental consonants (e.g. plosive, fricative or affricative phonemes – voiced or unvoiced, respectively), and other statistically detected phonetic characteristics of an association between position and movement of the lips and the voice and pronunciation of a speaker S_i.

The aforementioned noise reduction circuit 106 comprises digital signal processing means 106a for calculating a discrete signal spectrum $S(k\cdot\Delta f)$ that corresponds to an analog-to-digital-converted version $s(nT)$ of the recorded audio sequence $s(t)$ by performing a Fast Fourier Transform (FFT), audio feature extraction and analyzing means 106b (e.g. an amplitude detector) for detecting acoustic-phonetic speech characteristics of a speaker's voice and pronunciation based on the recorded audio sequence $s(t)$, means 106c for estimating the noise power density spectrum $\Phi_{nn}(f)$ of the statistically distributed background noise $n'(t)$ based on the result of the speaker detection procedure performed by said audio feature extraction and analyzing means 106b, a subtracting element 106d for subtracting a discretized version $\widetilde{\Phi}_{nn}(k\cdot\Delta f)$ of the estimated noise power density spectrum $\widetilde{\Phi}_{nn}(f)$ from the discrete signal spectrum $S(k\cdot\Delta f)$ of the analog-to-digital-converted audio sequence $s(nT)$, and digital signal processing means 106e for calculating the corresponding discrete time-domain signal $\hat{s}_i(nT)$ of the obtained difference signal by performing an Inverse Fast Fourier Transform (IFFT).

The depicted noise reduction and speech activity recognition system 100 comprises audio feature extraction and analyzing means 106b which are used for determining acoustic-phonetic speech characteristics of the speaker's voice and pronunciation ($\underline{o}_{a,nT}$) based on the recorded audio sequence $s(t)$ and visual feature extraction and analyzing means 104a+b for determining the current location of the speaker's face at a data rate of 1 frame/s, tracking lip movements and/or facial expressions of said speaker S_i at a data rate of 15 frames/s and determining acoustic-phonetic speech characteristics of the speaker's voice and pronunciation based on detected lip movements and/or facial expressions ($\underline{o}_{v,nT}$).

As depicted in Fig. 1, said noise reduction system 200b/c can advantageously be used for a video-telephony based application in a telecommunication system running on a video-enabled phone 102 which is equipped with a built-in video camera 101b' pointing at the face of a speaker S_i participating in a video telephony session.

Fig. 2b shows an example of a slow camera-enhanced noise reduction and speech activity recognition system 200b for a telephony-based application which implements an audio-

visual speech activity estimation algorithm according to one embodiment of the present invention. Thereby, an audio speech activity estimate taken from an audio feature vector $\underline{o}_{a,t}$ supplied by said audio feature extraction and analyzing means 106b is correlated with a further speech activity estimate that is obtained by calculating the difference of the discrete signal spectrum $S(k\cdot\Delta f)$ and a sampled version $\widetilde{\Phi}_{nn}(k\cdot\Delta f)$ of the estimated noise power density spectrum $\widetilde{\Phi}_{nn}(f)$ of the statistically distributed background noise $n'(t)$. Said audio speech activity estimate is obtained by an amplitude detection of the band-pass-filtered discrete signal spectrum $S(k\cdot\Delta f)$ of the analog-to-digital-converted audio signal $s(t)$.

Similar to the embodiment depicted in Fig. 1, the noise reduction and speech activity recognition system 200b depicted in Fig. 2b comprises an audio feature extraction and analyzing means 106b (e.g. an amplitude detector) which is used for determining acoustic-phonetic speech characteristics of the speaker's voice and pronunciation ($\underline{o}_{a,nT}$) based on the recorded audio sequence $s(t)$ and visual feature extraction and analyzing means 104' and 104'' for determining the current location of the speaker's face at a data rate of 1 frame/s, tracking lip movements and facial expressions of said speaker S_i at a data rate of 15 frames/s and determining acoustic-phonetic speech characteristics of the speaker's voice and pronunciation based on detected lip movements and/or facial expressions ($\underline{o}_{v,nT}$). Thereby, said audio feature extraction and analyzing means 106b can simply be realized as an amplitude detector.

Aside from the components 106a-e described above with reference to Fig. 1, the noise reduction circuit 106 depicted in Fig. 2b comprises a delay element 204, which provides a delayed version of the discrete signal spectrum $S(k\cdot\Delta f)$ of the analog-to-digital-converted audio signal $s(t)$, a first multiplier element 107a, which is used for correlating (S9) the discrete signal spectrum $S_\tau(k\cdot\Delta f)$ of a delayed version $s(nT\text{-}\tau)$ of the analog-to-digital-converted audio signal $s(nT)$ with a visual speech activity estimate taken from a visual feature vector $\underline{o}_{v,t}$ supplied by the visual feature extraction and analyzing means 104a+b and/or 104'+104'', thus yielding a further estimate $\widetilde{S}_i'(f)$ for updating the estimate $\widetilde{S}_i(f)$ for the frequency spectrum $S_i(f)$ corresponding to the signal $s_i(t)$ that represents said speaker's voice as well as a further estimate $\widetilde{\Phi}_{nn}'(f)$ for updating the estimate $\widetilde{\Phi}_{nn}(f)$ for the noise

power density spectrum $\Phi_{nn}(f)$ of the statistically distributed background noise $n'(t)$, and a second multiplier element 107, which is used for correlating (S8a) the discrete signal spectrum $S_\tau(k\cdot\Delta f)$ of a delayed version $s(nT-\tau)$ of the analog-to-digital-converted audio signal $s(nT)$ with an audio speech activity estimate obtained by an amplitude detection (S8b) of the band-pass-filtered discrete signal spectrum $S(k\cdot\Delta f)$, thus yielding an estimate $\widetilde{S}_i(f)$ for the frequency spectrum $S_i(f)$ which corresponds to the signal $s_i(t)$ that represents said speaker's voice and an estimate $\widetilde{\Phi}_{nn}(f)$ for the noise power density spectrum $\Phi_{nn}(f)$ of said background noise $n'(t)$. A sample-and-hold (S&H) element 106d' provides a sampled version $\widetilde{\Phi}_{nn}(k\cdot\Delta f)$ of the estimated noise power density spectrum $\widetilde{\Phi}_{nn}(f)$. The noise reduction circuit 106 further comprises a band-pass filter with adjustable cut-off frequencies, which is used for filtering the discrete signal spectrum $S(k\cdot\Delta f)$ of the analog-to-digital-converted audio signal $s(t)$. The cut-off frequencies can be adjusted dependent on the bandwidth of the estimated speech signal spectrum $\widetilde{S}_i(f)$. A switch 106f is provided for selectively switching between a first and a second mode for receiving said speech signal $s_i(t)$ with and without using the proposed audio-visual speech recognition approach providing a noise-reduced speech signal $\hat{s}_i(t)$, respectively. According to a further aspect of the present invention, means are provided for switching said microphone 101a off when the actual level of the speech activity indication signal $\hat{s}_i(nT)$ falls below a predefined threshold value (not shown).

An example of a fast camera-enhanced noise reduction and speech activity recognition system 200c for a telephony-based application which implements an audio-visual speech activity estimation algorithm according to a further embodiment of the present invention is depicted in Fig. 2c. The circuitry correlates a discrete signal spectrum $S(k\cdot\Delta f)$ of the analog-to-digital-converted audio signal $s(t)$ with a delayed version of an audio-visual speech activity estimate and a further speech activity estimate obtained by calculating the difference spectrum of the discrete signal spectrum $S(k\cdot\Delta f)$ and a sampled version $\widetilde{\Phi}_{nn}(k\cdot\Delta f)$ of the estimated noise power density spectrum $\widetilde{\Phi}_{nn}(f)$. The aforementioned audio-visual speech activity estimate is taken from an audio-visual feature vector $\underline{o}_{av,t}$ obtained by combining an audio feature vector $\underline{o}_{a,t}$ supplied by said audio feature extraction and analyzing means

106b with a visual feature vector $\underline{o}_{v,t}$ supplied by said visual speech activity detection module 104''.

Aside from the components described above with reference to Fig. 1, the noise reduction circuit 106 depicted in Fig. 2c comprises a summation element 107c, which is used for adding (S11a) an audio speech activity estimate supplied from an audio feature extraction and analyzing means 106b (e.g. an amplitude detector) for determining acoustic-phonetic speech characteristics of the speaker's voice and pronunciation ($\underline{o}_{a,nT}$) based on the recorded audio sequence $s(t)$ to an visual speech activity estimate supplied from visual feature extraction and analyzing means 104' and 104'' for determining the current location of the speaker's face at a data rate of 1 frame/s, tracking lip movements and facial expressions of said speaker S_i at a data rate of 15 frames/s and determining acoustic-phonetic speech characteristics of the speaker's voice and pronunciation based on detected lip movements and/or facial expressions ($\underline{o}_{v,nT}$), thus yielding an audio-visual speech activity estimate. The noise reduction circuit 106 further comprises a multiplier element 107', which is used for correlating (S11b) the discrete signal spectrum $S(k\cdot\Delta f)$ of the analog-to-digital-converted audio signal $s(t)$ with an audio-visual speech activity estimate, obtained by combining an audio feature vector $\underline{o}_{a,t}$ supplied by said audio feature extraction and analyzing means 106b with a visual feature vector $\underline{o}_{v,t}$ supplied by said visual speech activity detection module 104'', thereby yielding an estimate $\widetilde{S}_i(f)$ for the frequency spectrum $S_i(f)$ which corresponds to the signal $s_i(t)$ that represents the speaker's voice and an estimate $\widetilde{\Phi}_{nn}(f)$ for the noise power density spectrum $\Phi_{nn}(f)$ of the statistically distributed background noise $n'(t)$. A sample-and-hold (S&H) element 106d' provides a sampled version $\widetilde{\Phi}_{nn}(k\cdot\Delta f)$ of the estimated noise power density spectrum $\widetilde{\Phi}_{nn}(f)$. The noise reduction circuit 106 further comprises a band-pass filter with adjustable cut-off frequencies, which is used for filtering the discrete signal spectrum $S(k\cdot\Delta f)$ of the analog-to-digital-converted audio signal $s(t)$. Said cut-off frequencies can be adjusted dependent on the bandwidth of the estimated speech signal spectrum $\widetilde{S}_i(f)$. A switch 106f is provided for selectively switching between a first and a second mode for receiving said speech signal $s_i(t)$ with and without using the proposed audio-visual speech recognition approach providing a noise-reduced

speech signal $\hat{s}_i(t)$, respectively. According to a further aspect of the present invention, said noise reduction system 200c comprises means (SW) for switching said microphone 101a off when the actual level of the speech activity indication signal $\hat{s}_i(nT)$ falls below a predefined threshold value (not shown).

A still further embodiment of the present invention is directed to a near-end speaker detection method as shown in the flow chart depicted in Fig. 3a. Said method reduces the noise level of a recorded analog audio sequence $s(t)$ being interfered by a statistically distributed background noise $n'(t)$, said audio sequence representing the voice of a speaker S_i. After having subjected (S1) the analog audio sequence $s(t)$ to an analog-to-digital conversion, the corresponding discrete signal spectrum $S(k\cdot\Delta f)$ of the analog-to-digital-converted audio sequence $s(nT)$ is calculated (S2) by performing a Fast Fourier Transform (FFT) and the voice of said speaker S_i is detected (S3) from said signal spectrum $S(k\cdot\Delta f)$ by analyzing visual features extracted from a simultaneously with the recording of the analog audio sequence $s(t)$ recorded video sequence $v(nT)$ tracking the current location of the speaker's face, lip movements and/or facial expressions of the speaker S_i in subsequent images. Next, the noise power density spectrum $\tilde{\Phi}_{nn}(f)$ of the statistically distributed background noise $n'(t)$ is estimated (S4) based on the result of the speaker detection step (S3), whereupon a sampled version $\tilde{\Phi}_{nn}(k\cdot\Delta f)$ of the estimated noise power density spectrum $\tilde{\Phi}_{nn}(f)$ is subtracted (S5) from the discrete spectrum $S(k\cdot\Delta f)$ of the analog-to-digital-converted audio sequence $s(nT)$. Finally, the corresponding discrete time-domain signal $\hat{s}_i(nT)$ of the obtained difference signal, which represents a discrete version of the recognized speech signal, is calculated (S6) by performing an Inverse Fast Fourier Transform (IFFT).

Optionally, a multi-channel acoustic echo cancellation algorithm which models echo path impulse responses by means of adaptive finite impulse response (FIR) filters and subtracts echo signals from the analog audio sequence $s(t)$ can be conducted (S7) based on acoustic-phonetic speech characteristics derived by an algorithm for extracting visual features from a video sequence tracking the location of a speaker's face, lip movements and/or facial expressions of the speaker S_i in subsequent images. Said multi-channel acoustic echo cancellation algorithm thereby performs a double-talk detection procedure.

According to a further aspect of the invention, a learning procedure is applied which enhances the step of detecting (S3) the voice of said speaker S_i from the discrete signal spectrum $S(k\cdot\Delta f)$ of the analog-to-digital-converted version $s(nT)$ of an analog audio sequence $s(t)$ by analyzing visual features extracted from a simultaneously with the recording of the analog audio sequence $s(t)$ recorded video sequence tracking the current location of the speaker's face, lip movements and/or facial expressions of the speaker S_i in subsequent images.

In one embodiment of the present invention, which is illustrated in the flow charts depicted in Figs. 3a+b, a near-end speaker detection method is proposed that is characterized by the step of correlating (S8a) the discrete signal spectrum $S_\tau(k\cdot\Delta f)$ of a delayed version $s(nT\text{-}\tau)$ of the analog-to-digital-converted audio signal $s(nT)$ with an audio speech activity estimate obtained by an amplitude detection (S8b) of the band-pass-filtered discrete signal spectrum $S(k\cdot\Delta f)$, thereby yielding an estimate $\widetilde{S}_i(f)$ for the frequency spectrum $S_i(f)$ which corresponds to the signal $s_i(t)$ representing said speaker's voice and an estimate $\widetilde{\Phi}_{nn}(f)$ for the noise power density spectrum $\Phi_{nn}(f)$ of said background noise $\tilde{n}(t)$. Moreover, the discrete signal spectrum $S_\tau(k\cdot\Delta f)$ of a delayed version $s(nT\text{-}\tau)$ of the analog-to-digital-converted audio signal $s(nT)$ is correlated (S9) with a visual speech activity estimate taken from a visual feature vector $\underline{o}_{v,t}$ which is supplied by the visual feature extraction and analyzing means 104a+b and/or 104'+104'', thus yielding a further estimate $\widetilde{S}_i'(f)$ for updating the estimate $\widetilde{S}_i(f)$ for the frequency spectrum $S_i(f)$ which corresponds to the signal $s_i(t)$ representing the speaker's voice as well as a further estimate $\widetilde{\Phi}_{nn}'(f)$ that is used for updating the estimate $\widetilde{\Phi}_{nn}(f)$ for the noise power density spectrum $\Phi_{nn}(f)$ of the statistically distributed background noise $n'(t)$. The noise reduction circuit 106 thereby provides a band-pass filter 204 for filtering the discrete signal spectrum $S(k\cdot\Delta f)$ of the analog-to-digital-converted audio signal $s(t)$, wherein the cut-off frequencies of said band-pass filter 204 are adjusted (S10) dependent on the bandwidth of the estimated speech signal spectrum $\widetilde{S}_i(f)$.

In a further embodiment of the present invention as shown in the flow charts depicted in Figs. 3a+c a near-end speaker detection method is proposed which is characterized by the step of adding (S11a) an audio speech activity estimate obtained by an amplitude detection of the band-pass-filtered discrete signal spectrum $S(k\cdot\Delta f)$ of the analog-to-digital-converted audio signal $s(t)$ to a visual speech activity estimate taken from a visual feature vector $\underline{o}_{v,t}$ supplied by said visual feature extraction and analyzing means 104a+b and/or 104'+104'', thereby yielding an audio-visual speech activity estimate. According to this embodiment, the discrete signal spectrum $S(k\cdot\Delta f)$ is correlated (S11b) with the audio-visual speech activity estimate, thus yielding an estimate $\widetilde{S}_t(f)$ for the frequency spectrum $S_t(f)$ corresponding to the signal $s_t(t)$ that represents said speaker's voice as well as an estimate $\widetilde{\Phi}_{nn}(f)$ for the noise power density spectrum $\Phi_{nn}(f)$ of the statistically distributed background noise $n'(t)$. The cut-off frequencies of the band-pass filter 204 that is used for filtering the discrete signal spectrum $S(k\cdot\Delta f)$ of the analog-to-digital-converted audio signal $s(t)$ are adjusted (S11c) dependent on the bandwidth of the estimated speech signal spectrum $\widetilde{S}_t(f)$.

Finally, the present invention also pertains to the use of a noise reduction system 200b/c and a corresponding near-end speaker detection method as described above for a video-telephony based application (e.g. a video conference) in a telecommunication system running on a video-enabled phone having a built-in video camera 101b' pointing at the face of a speaker S_t participating in a video telephony session. This especially pertains to a scenario where a number of persons are sitting in one room equipped with many cameras and microphones such that a speaker's voice interferes with the voices of the other persons.

Table: Depicted Features and Their Corresponding Reference Signs

No.	Technical Feature (System Component or Procedure Step)
100	noise reduction and speech activity recognition system having an audio-visual user interface, said system being specially adapted for running a real-time lip tracking application which combines visual features $o_{v,nT}$ extracted from a digital video sequence $v(nT)$ showing the face of a speaker S_i by detecting and analyzing the speaker's lip movements and/or facial expressions with audio features $o_{a,nT}$ extracted from an analog audio sequence $s(t)$ representing the voice of said speaker S_i interfered by a statistically distributed background noise $n'(t)$, wherein said audio sequence $s(t)$ includes – aside from the signal representing the voice of said speaker S_i – both environmental noise $n(t)$ and a weighted sum $\sum_j a_j \cdot s_j(t-T_j)$ $(j \neq i)$ of surrounding persons' interfering voices in the environment of said speaker S_i
101a	microphone, used for recording an analog audio sequence $s(t)$ representing the voice of a speaker S_i interfered by a statistically distributed background noise $n'(t)$, which includes both environmental noise $n(t)$ and a weighted sum $\sum_j a_j \cdot s_j(t-T_j)$ (with $j \neq i$) of surrounding persons' interfering voices in the environment of said speaker S_i
101a'	analog-to-digital converter (ADC), used for converting the analog audio sequence $s(t)$ recorded by said microphone 101a into the digital domain
101b	video camera pointing to the speaker's face for recording a video sequence showing lip movements and/or facial expressions of said speaker S_i
101b'	video camera as described above with an integrated analog-to-digital converter (ADC)
102	video telephony application, used for transmitting a video sequence showing a speaker's face and lip movements in subsequent images
104	visual front end of an automatic audio-visual speech recognition system 100 using a bimodal approach to speech recognition and near-speaker detection by incorporating a real-time lip tracking algorithm for deriving additional visual features from lip move-

No.	Technical Feature (System Component or Procedure Step)
	ments and/or facial expressions of a speaker S_i whose voice is interfered by a statistically distributed background noise $n'(t)$, the visual front end 104 comprising visual feature extraction and analyzing means for continuously or intermittently determining the current location of the speaker's face, tracking lip movements and/or facial expressions of the speaker S_i in subsequent images and determining acoustic-phonetic speech characteristics of the speaker's voice and pronunciation based on detected lip movements and/or facial expressions
104'	visual feature extraction module for continuously tracking lip movements and/or facial expressions of the speaker S_i and determining acoustic-phonetic speech characteristics of the speaker's voice based on detected lip movements and/or facial expressions
104''	visual speech activity detection module for analyzing the acoustic-phonetic speech characteristics and detecting speech activity of a speaker based on said analysis
104a	visual feature extraction means for continuously or intermittently determining the current location of the speaker's face recorded by a video camera 101b at a rate of 1 frame/s
104b	visual feature extraction and analyzing means for continuously tracking lip movements and/or facial expressions of the speaker S_i and determining acoustic-phonetic speech characteristics of said speaker's voice based on detected lip movements and/or facial expressions at a rate of 15 frames/s
106	noise reduction circuit being specially adapted to reduce statistically distributed background noise $n'(t)$ received by said microphone 101a and perform a near-speaker detection by separating the speaker's voice from said background noise $n'(t)$ based on a combination of the speech characteristics which are derived by said audio and visual feature extraction and analyzing means 104a+b and 106b, respectively
106a	digital signal processing means for calculating the discrete signal spectrum $S(k \cdot \Delta f)$ that corresponds to an analog-to-digital-converted version $s(nT)$ of the recorded audio sequence $s(t)$ by performing a Fast Fourier Transform (FFT)
106b	audio feature extraction and analyzing means (e.g. an amplitude detector) for detecting acoustic-phonetic speech characteristics of the speaker's voice and pronunciation based on the recorded audio sequence $s(t)$
106c	means for estimating the noise power density spectrum $\Phi_{nn}(f)$ of the statistically dis-

No.	Technical Feature (System Component or Procedure Step)
	tributed background noise $n'(t)$ based on the result of the speaker detection procedure performed by said audio-visual feature extraction and analyzing means 104b, 106b, 104' and/or 104''
106c'	means for estimating the signal spectrum $S_i(f)$ of the recorded speech signal $s_i(t)$ based on the result of the speaker detection procedure performed by said audio-visual feature extraction and analyzing means 104b, 106b, 104' and/or 104''
106d	subtracting element for subtracting a discretized version $\widetilde{\Phi}_{nn}(k \cdot \Delta f)$ of the estimated noise power density spectrum $\widetilde{\Phi}_{nn}(f)$ from the discrete signal spectrum $S(k \cdot \Delta f)$ of the analog-to-digital-converted audio sequence $s(nT)$
106d'	sample-and-hold (S&H) element providing a sampled version $\widetilde{\Phi}_{nn}(k \cdot \Delta f)$ of the estimated noise power density spectrum $\widetilde{\Phi}_{nn}(f)$
106e	digital signal processing means for calculating the corresponding discrete time-domain signal $\hat{s}_i(nT)$ of the obtained difference signal by performing an Inverse Fast Fourier Transform (IFFT)
106f	switch for selectively switching between a first and a second mode for receiving said speech signal $s_i(t)$ with and without using the proposed audio-visual speech recognition approach providing a noise-reduced speech signal $\hat{s}_i(t)$, respectively
107	multiplier element, used for correlating the discrete signal spectrum $S(k \cdot \Delta f)$ of the analog-to-digital-converted audio signal $s(t)$ with an audio speech activity estimate which is obtained by an amplitude detection of the digital audio signal $s(nT)$
107'	multiplier element, used for correlating (S11b) the discrete signal spectrum $S(k \cdot \Delta f)$ of the analog-to-digital-converted audio signal $s(t)$ with an audio-visual speech activity estimate, obtained by combining an audio feature vector $\underline{o}_{a,t}$ supplied by said audio feature extraction and analyzing means 106b with a visual feature vector $\underline{o}_{v,t}$ supplied by said visual speech activity detection module 104'', thereby yielding an estimate $\widetilde{S}_i(f)$ for the frequency spectrum $S_i(f)$ corresponding to the signal $s_i(t)$ which represents said speaker's voice and an estimate $\widetilde{\Phi}_{nn}(f)$ for the noise power density spectrum $\Phi_{nn}(f)$ of the statistically distributed background noise $n'(t)$

No.	Technical Feature (System Component or Procedure Step)
107a	multiplier element, used for correlating (S9) the discrete signal spectrum $S_\tau(k\cdot\Delta f)$ of a delayed version $s(nT-\tau)$ of the analog-to-digital-converted audio signal $s(nT)$ with a visual speech activity estimate taken from a visual feature vector $\underline{o}_{v,t}$ supplied by the visual feature extraction and analyzing means 104a+b and/or 104'+104'', thereby yielding a further estimate $\widetilde{S}_i'(f)$ for updating the estimate $\widetilde{S}_i(f)$ for the frequency spectrum $S_i(f)$ corresponding to the signal $s_i(t)$ which represents said speaker's voice as well as a further estimate $\widetilde{\Phi}_{nn}'(f)$ for updating the estimate $\widetilde{\Phi}_{nn}(f)$ for the noise power density spectrum $\Phi_{nn}(f)$ of the statistically distributed background noise $n'(t)$
107b	multiplier element, used for correlating (S8a) the discrete signal spectrum $S_\tau(k\cdot\Delta f)$ of a delayed version $s(nT-\tau)$ of the analog-to-digital-converted audio signal $s(nT)$ with an audio speech activity estimate obtained by an amplitude detection (S8b) of the band-pass-filtered discrete signal spectrum $S(k\cdot\Delta f)$, thereby yielding an estimate $\widetilde{S}_i(f)$ for the frequency spectrum $S_i(f)$ corresponding to the signal $s_i(t)$ which represents said speaker's voice as well as an estimate $\widetilde{\Phi}_{nn}(f)$ for the noise power density spectrum $\Phi_{nn}(f)$ of the statistically distributed background noise $n'(t)$
107c	summation element, used for adding (S11a) the audio speech activity estimate to the visual speech activity estimate, thereby yielding an audio-visual speech activity estimate
108	multi-channel acoustic echo cancellation unit being specially adapted to perform a near-end speech detection and/or double-talk detection algorithm based on acoustic-phonetic speech characteristics derived by said audio and visual feature extraction and analyzing means 104a+b and 106b, respectively
108a	means for near-end talk and/or double-talk detection, integrated in the multi-channel acoustic echo cancellation unit 108
200a	block diagram showing a conventional noise reduction and speech activity recognition system for a telephony-based application based on an audio speech activity estimation according to the state of the art, wherein the discrete signal spectrum $S(k\cdot\Delta f)$ of the analog-to-digital-converted audio signal $s(t)$ is correlated with an audio speech activity estimate which is obtained by an amplitude detection of the digital audio signal $s(nT)$
200b	block diagram showing an example of a slow camera-enhanced noise reduction and

No.	Technical Feature (System Component or Procedure Step)
	speech activity recognition system for a telephony-based application implementing an audio-visual speech activity estimation algorithm according to one embodiment of the present invention, wherein the discrete signal spectrum $S_\tau(k\cdot\Delta f)$ of a delayed version $s(nT-\tau)$ of the analog-to-digital-converted audio signal $s(nT)$ is correlated (S8a) with an audio speech activity estimate obtained by an amplitude detection (S8b) of the band-pass-filtered discrete signal spectrum $S(k\cdot\Delta f)$, thereby yielding an estimate $\widetilde{S}_i(f)$ for the frequency spectrum $S_i(f)$ corresponding to the signal $s_i(t)$ which represents said speaker's voice and an estimate $\widetilde{\Phi}_{nn}(f)$ for the noise power density spectrum $\Phi_{nn}(f)$ of the statistically distributed background noise $n'(t)$, and also correlated (S9) with a visual speech activity estimate taken from a visual feature vector $\underline{o}_{v,t}$ supplied by the visual feature extraction and analyzing means 104a+b and/or 104'+104'', thereby yielding a further estimate $\widetilde{S}_i'(f)$ for updating the estimate $\widetilde{S}_i(f)$ for the frequency spectrum $S_i(f)$ corresponding to the signal $s_i(t)$ which represents said speaker's voice as well as a further estimate $\widetilde{\Phi}_{nn}'(f)$ for updating the estimate $\widetilde{\Phi}_{nn}(f)$ for the noise power density spectrum $\Phi_{nn}(f)$ of the statistically distributed background noise $n'(t)$
200c	block diagram showing an example of a fast camera-enhanced noise reduction and speech activity recognition system for a telephony-based application implementing an audio-visual speech activity estimation algorithm according to a further embodiment of the present invention, wherein the discrete signal spectrum $S(k\cdot\Delta f)$ of the analog-to-digital-converted audio signal $s(t)$ is correlated (S11b) with an audio-visual speech activity estimate, obtained by combining an audio feature vector $\underline{o}_{a,t}$ which is supplied by said audio feature extraction and analyzing means 106b with a visual feature vector $\underline{o}_{v,t}$ supplied by the visual speech activity detection module 104'', thereby yielding an estimate $\widetilde{S}_i(f)$ for the corresponding frequency spectrum $S_i(f)$ of the signal $s_i(t)$ which represents said speaker's voice as well as an estimate $\widetilde{\Phi}_{nn}(f)$ for the noise power density spectrum $\Phi_{nn}(f)$ of the statistically distributed background noise $n'(t)$
202	delay element, providing a delayed version of the discrete signal spectrum $S(k\cdot\Delta f)$ of the analog-to-digital-converted audio signal $s(t)$

No.	Technical Feature (System Component or Procedure Step)
204	band-pass filter with adjustable cut-off frequencies which can be adjusted dependent on the bandwidth of the estimated speech signal spectrum $\widetilde{S}_t(f)$, used for filtering the discrete signal spectrum $S(k\cdot\Delta f)$ of the analog-to-digital-converted audio signal $s(t)$
300a	flow chart illustrating a near-end speaker detection method reducing the noise level of a detected analog audio sequence $s(t)$ according to the embodiment depicted in Fig. 1 of the present invention
300b	flow chart illustrating a near-end speaker detection method according to the embodiment depicted in Fig. 2b of the present invention
300c	flow chart illustrating a near-end speaker detection method according to the embodiment depicted in Fig. 2c of the present invention
SW	means for switching said microphone 101a off when the actual level of the speech activity indication signal $\hat{s}_t(nT)$ falls below a predefined threshold value (not shown)
S1	step #1: subjecting the analog audio sequence $s(t)$ to an analog-to-digital conversion
S10	step #10: adjusting the cut-off frequencies of the band-pass filter 204 used for filtering the discrete signal spectrum $S(k\cdot\Delta f)$ of the analog-to-digital-converted audio signal ($s(t)$) dependent on the bandwidth of the estimated speech signal spectrum $\widetilde{S}_t(f)$
S11a	step #11a: adding an audio speech activity estimate which is obtained by an amplitude detection of the band-pass-filtered discrete signal spectrum $S(k\cdot\Delta f)$ of the analog-to-digital-converted audio signal $s(t)$ to a visual speech activity estimate taken from a visual feature vector $\underline{o}_{v,t}$ supplied by the visual feature extraction and analyzing means 104a+b and/or 104'+ 104'', thereby yielding an audio-visual speech activity estimate
S11b	step #11b: correlating the discrete signal spectrum $S(k\cdot\Delta f)$ with the audio-visual speech activity estimate, thereby yielding an estimate $\widetilde{S}_t(f)$ for the frequency spectrum $S_t(f)$ corresponding to the signal $s_t(t)$ which represents said speaker's voice as well as an estimate $\widetilde{\Phi}_{nn}(f)$ for said noise power density spectrum $\Phi_{nn}(f)$
S11c	step #11c: adjusting the cut-off frequencies of a band-pass filter 204 used for filtering the discrete signal spectrum $S(k\cdot\Delta f)$ of the analog-to-digital-converted audio signal $s(t)$ dependent on the bandwidth of the estimated speech signal spectrum $\widetilde{S}_t(f)$

No.	Technical Feature (System Component or Procedure Step)
S2	step #2: calculating the corresponding discrete signal spectrum $S(k\cdot\Delta f)$ of the analog-to-digital-converted audio sequence $s(nT)$ by performing a Fast Fourier Transform (FFT)
S3	step #3: detecting the voice of said speaker S_i from said signal spectrum $S(k\cdot\Delta f)$ by analyzing visual features extracted from a simultaneously with the recording of the audio sequence $s(t)$ recorded video sequence for tracking the current location of the speaker's face, lip movements and/or facial expressions of the speaker S_i in subsequent images,
S4	step #4: estimating the noise power density spectrum $\Phi_{nn}(f)$ of the statistically distributed background noise $n'(t)$ based on the result of the speaker detection step S3
S5	step #5: subtracting a discretized version $\widetilde{\Phi}_{nn}(k\cdot\Delta f)$ of the estimated noise power density spectrum $\widetilde{\Phi}_{nn}(f)$ from the discrete signal spectrum $S(k\cdot\Delta f)$ of the analog-to-digital-converted audio sequence $s(nT)$
S6	step #6: calculating the corresponding discrete time-domain signal $\hat{s}_i(nT)$ of the obtained difference signal by performing an Inverse Fast Fourier Transform (IFFT)
S7	step #7: conducting a multi-channel acoustic echo cancellation algorithm which models echo path impulse responses by means of adaptive finite impulse response (FIR) filters and subtracts echo signals from the analog audio sequence $s(t)$ based on acoustic-phonetic speech characteristics derived by an algorithm for extracting visual features from a video sequence tracking the location of a speaker's face, lip movements and/or facial expressions of the speaker S_i in subsequent images
S8o	step #8o: band-pass-filtering the discrete signal spectrum $S(k\cdot\Delta f)$ of the analog-to-digital-converted audio signal $s(nT)$
S8a	step #8a: correlating the discrete signal spectrum $S_\tau(k\cdot\Delta f)$ of a delayed version $s(nT-\tau)$ of the analog-to-digital-converted audio signal $s(nT)$ with an audio speech activity estimate obtained by the amplitude detection step S8b
S8b	step #8b: amplitude detection of the band-pass-filtered discrete signal spectrum $S(k\cdot\Delta f)$, thereby yielding an estimate $\widetilde{S}_i(f)$ for the frequency spectrum $S_i(f)$ corresponding to the signal $s_i(t)$ which represents said speaker's voice as well as an estimate $\widetilde{\Phi}_{nn}(f)$ for the noise power density spectrum $\Phi_{nn}(f)$

No.	Technical Feature (System Component or Procedure Step)
S9	step #9: correlating the discrete signal spectrum $S_\tau(k\cdot\Delta f)$ of a delayed version $s(nT-\tau)$ of the analog-to-digital-converted audio signal $s(nT)$ with a visual speech activity estimate taken from a visual feature vector $\underline{o}_{v,t}$ supplied by the visual feature extraction and analyzing means 104a+b and/or 104'+104'', thereby yielding a further estimate $\widetilde{S}_i'(f)$ for updating the estimate $\widetilde{S}_i(f)$ for the frequency spectrum $S_i(f)$ corresponding to the signal $s_i(t)$ which represents said speaker's voice as well as a further estimate $\widetilde{\Phi}_{nn}'(f)$ for updating the estimate $\widetilde{\Phi}_{nn}(f)$ for the noise power density spectrum $\Phi_{nn}(f)$ of the statistically distributed background noise $n'(t)$
S10	step #10: adjusting the cut-off frequencies of a band-pass filter 204 used for filtering the discrete signal spectrum $S(k\cdot\Delta f)$ of the analog-to-digital-converted audio signal $s(t)$ dependent on the bandwidth of the estimated speech signal spectrum $\widetilde{S}_i(f)$
S11a	step #11a: adding an audio speech activity estimate obtained by an amplitude detection of the band-pass-filtered discrete signal spectrum $S(k\cdot\Delta f)$ of the analog-to-digital-converted audio signal $s(t)$ to a visual speech activity estimate taken from a visual feature vector $\underline{o}_{v,t}$ supplied by said visual feature extraction and analyzing means 104a+b, and/or 104'+104'', thereby yielding an audio-visual speech activity estimate
S11b	step #11b: correlating the discrete signal spectrum $S(k\cdot\Delta f)$ with the audio-visual speech activity estimate, thus yielding an estimate $\widetilde{S}_i(f)$ for the frequency spectrum $S_i(f)$ corresponding to the signal $s_i(t)$ which represents said speaker's voice as well as an estimate $\widetilde{\Phi}_{nn}(f)$ for the noise power density spectrum $\Phi_{nn}(f)$ of the statistically distributed background noise $n'(t)$
S11c	step #11c: adjusting the cut-off frequencies of a band-pass filter 204 used for filtering the discrete signal spectrum $S(k\cdot\Delta f)$ of the analog-to-digital-converted audio signal $s(t)$ dependent on the bandwidth of the estimated speech signal spectrum $\widetilde{S}_i(f)$

Claims

1. A noise reduction system with an audio-visual user interface, said system being specially adapted for running an application for combining visual features ($o_{v,nT}$) extracted from a digital video sequence ($v(nT)$) showing the face of a speaker (S_i) with audio features ($o_{a,nT}$) extracted from an analog audio sequence ($s(t)$), wherein said audio sequence ($s(t)$) can include noise in the environment of said speaker (S_i), said noise reduction system (200b/c) comprising

- means (101a, 106b) for detecting and analyzing said analog audio sequence ($s(t)$),
- means (101b') for detecting said video sequence ($v(nT)$), and
- means (104a+b, 104'+104'') for analyzing the detected video signal ($v(nT)$),

characterized by

a noise reduction circuit (106) being adapted to separate the speaker's voice from said background noise ($n'(t)$) based on a combination of derived speech characteristics ($o_{av,nT} := [o_{a,nT}^T, o_{v,nT}^T]^T$) and outputting a speech activity indication signal ($\hat{s}_i(nT)$) which is obtained by a combination of speech activity estimates supplied by said analyzing means (106b, 104a+b, 104'+104'').

2. A noise reduction system according to claim 1,

characterized by

means (SW) for switching off an audio channel in case the actual level of said speech activity indication signal ($\hat{s}_i(nT)$) falls below a predefined threshold value.

3. A noise reduction system according to anyone of the claims 1 or 2,

characterized by

a multi-channel acoustic echo cancellation unit (108) being specially adapted to perform a near-end speaker detection and double-talk detection algorithm based on acoustic-phonetic speech characteristics derived by said audio feature extraction and analyzing means (106b) and said visual feature extraction and analyzing means (104a+b, 104'+104'').

4. A noise reduction system according to anyone of the claims 1 to 3, characterized in that said audio feature extraction and analyzing means (106b) is an amplitude detector.

5. A near-end speaker detection method reducing the noise level of a detected analog audio sequence ($s(t)$), said method being characterized by the following steps:

- subjecting (S1) said analog audio sequence ($s(t)$) to an analog-to-digital conversion,
- calculating (S2) the corresponding discrete signal spectrum ($S(k\cdot\Delta f)$) of the analog-to-digital-converted audio sequence ($s(nT)$) by performing a Fast Fourier Transform (FFT),
- detecting (S3) the voice of said speaker (S_i) from said signal spectrum ($S(k\cdot\Delta f)$) by analyzing visual features ($o_{v,nT}$) extracted from a simultaneously with the recording of the analog audio sequence ($s(t)$) recorded video sequence ($v(nT)$) tracking the current location of the speaker's face, lip movements and/or facial expressions of the speaker (S_i) in subsequent images,
- estimating (S4) the noise power density spectrum ($\Phi_{nn}(f)$) of the statistically distributed background noise ($\tilde{n}(t)$) based on the result of the speaker detection step (S3),
- subtracting (S5) a discretized version ($\tilde{\Phi}_{nn}(k\cdot\Delta f)$) of the estimated noise power density spectrum ($\tilde{\Phi}_{nn}(f)$) from the discrete signal spectrum ($S(k\cdot\Delta f)$) of the analog-to-digital-converted audio sequence ($s(nT)$), and
- calculating (S6) the corresponding discrete time-domain signal ($\hat{s}_i(nT)$) of the obtained difference signal by performing an Inverse Fast Fourier Transform (IFFT), thereby yielding a discrete version of the recognized speech signal.

6. A near-end speaker detection method according to claim 5, characterized by the step of conducting (S7) a multi-channel acoustic echo cancellation algorithm which models echo path impulse responses by means of adaptive finite impulse response (FIR) filters and subtracts echo signals from the analog audio sequence ($s(t)$) based on acoustic-phonetic speech characteristics derived by an algorithm for extracting visual features ($o_{v,nT}$) from a video

sequence ($v(nT)$) tracking the location of a speaker's face, lip movements and/or facial expressions of the speaker (S_i) in subsequent images.

7. A near-end speaker detection method according to claim 6,
characterized in that
said multi-channel acoustic echo cancellation algorithm performs a double-talk detection procedure.

8. A near-end speaker detection method according to anyone of the claims 5 to 7,
characterized in that
said acoustic-phonetic speech characteristics are based on the opening of a speaker's mouth as an estimate of the acoustic energy of articulated vowels or diphthongs, respectively, rapid movement of the speaker's lips as a hint to labial or labio-dental consonants, respectively, and other statistically detected phonetic characteristics of an association between position and movement of the lips and the voice and pronunciation of said speaker (S_i).

9. A near-end speaker detection method according to anyone of the claims 5 to 8,
characterized by
a learning procedure used for enhancing the step of detecting (S3) the voice of said speaker (S_i) from the discrete signal spectrum ($S(k\cdot\Delta f)$) of the analog-to-digital-converted version ($s(nT)$) of an analog audio sequence ($s(t)$) by analyzing visual features ($o_{v,nT}$) extracted from a simultaneously with the recording of the analog audio sequence ($s(t)$) recorded video sequence ($v(nT)$) tracking the current location of the speaker's face, lip movements and/or facial expressions of the speaker (S_i) in subsequent images.

10. A near-end speaker detection method according to anyone of the claims 5 to 9,
characterized by the step of
correlating (S8a) the discrete signal spectrum ($S_\tau(k\cdot\Delta f)$) of a delayed version ($s(nT-\tau)$) of the analog-to-digital-converted audio signal ($s(nT)$) with an audio speech activity estimate obtained by an amplitude detection (S8b) of the band-pass-filtered discrete signal spectrum ($S(k\cdot\Delta f)$), thereby yielding an estimate ($\widetilde{S}_i(f)$) for the frequency spectrum ($S_i(f)$) corre-

sponding to the signal ($s_i(t)$) which represents said speaker's voice as well as an estimate ($\widetilde{\Phi}_{nn}(f)$) for the noise power density spectrum ($\Phi_{nn}(f)$) of the statistically distributed background noise ($n'(t)$).

11. A near-end speaker detection method according to claim 10,
characterized by the step of
correlating (S9) the discrete signal spectrum ($S_\tau(k\cdot\Delta f)$) of a delayed version ($s(nT-\tau)$) of the analog-to-digital-converted audio signal ($s(nT)$) with a visual speech activity estimate taken from a visual feature vector ($\underline{o}_{v,t}$) supplied by the visual feature extraction and analyzing means (104a+b, 104'+104''), thereby yielding a further estimate ($\widetilde{S}_i'(f)$) for updating the estimate ($\widetilde{S}_i(f)$) for the frequency spectrum ($S_i(f)$) corresponding to the signal ($s_i(t)$) which represents said speaker's voice as well as a further estimate ($\widetilde{\Phi}_{nn}'(f)$) for updating the estimate ($\widetilde{\Phi}_{nn}(f)$) for the noise power density spectrum ($\Phi_{nn}(f)$) of the statistically distributed background noise ($n'(t)$).

12. A near-end speaker detection method according anyone of the claims 10 or 11,
characterized by the step of
adjusting (S10) the cut-off frequencies of a band-pass filter (204) used for filtering the discrete signal spectrum ($S(k\cdot\Delta f)$) of the analog-to-digital-converted audio signal ($s(t)$) dependent on the bandwidth of the estimated speech signal spectrum ($\widetilde{S}_i(f)$).

13. A near-end speaker detection method according to anyone of the claims 5 to 9,
characterized by the steps of
- adding (S11a) an audio speech activity estimate obtained by an amplitude detection of the band-pass-filtered discrete signal spectrum ($S(k\cdot\Delta f)$) of the analog-to-digital-converted audio signal ($s(t)$) to a visual speech activity estimate taken from a visual feature vector ($\underline{o}_{v,t}$) supplied by said visual feature extraction and analyzing means (104a+b, 104'+104''), thereby yielding an audio-visual speech activity estimate,
- correlating (S11b) the discrete signal spectrum ($S(k\cdot\Delta f)$) with the audio-visual speech activity estimate, thereby yielding an estimate ($\widetilde{S}_i(f)$) for the frequency spectrum ($S_i(f)$)

corresponding to the signal ($s_i(t)$) which represents said speaker's voice as well as an estimate ($\widetilde{\Phi}_{nn}(f)$) for the noise power density spectrum ($\Phi_{nn}(f)$) of the statistically distributed background noise ($n'(t)$) and

- adjusting (S11c) the cut-off frequencies of a band-pass filter (204) used for filtering the discrete signal spectrum ($S(k\cdot\Delta f)$) of the analog-to-digital-converted audio signal ($s(t)$) dependent on the bandwidth of the estimated speech signal spectrum ($\widetilde{S}_i(f)$).

14. Use of a noise reduction system (200b/c) according to anyone of the claims 1 to 4 and a near-end speaker detection method according to anyone of the claims 5 to 13 for a video-telephony based application in a telecommunication system running on a video-enabled phone with a built-in video camera (101b') pointing at the face of a speaker (S_i) participating in a video telephony session.

15. A telecommunication device equipped with an audio-visual user interface, characterized by
noise reduction system (200b/c) according to anyone of the claims 1 to 4.

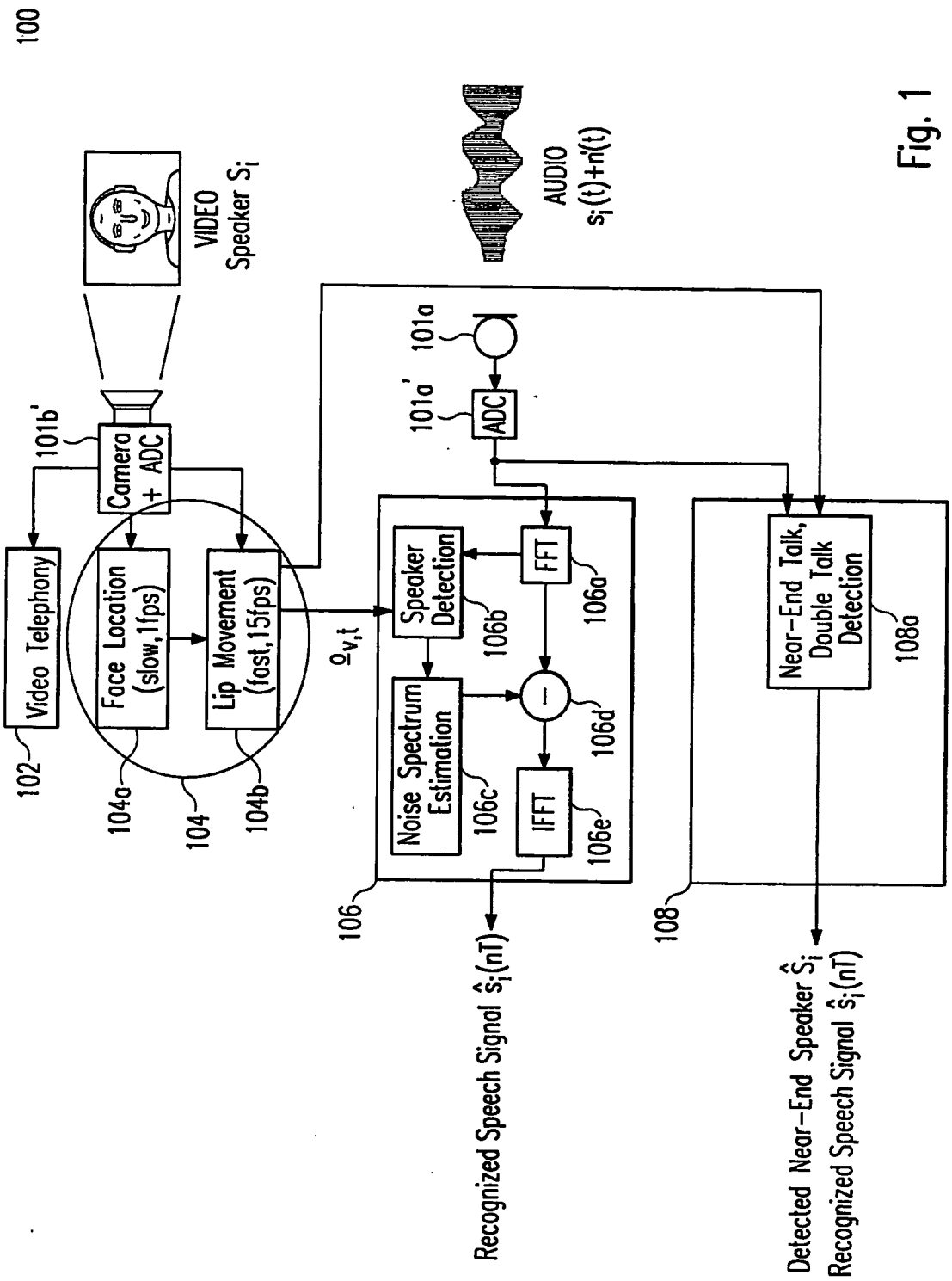

100
VIDEO
Speaker S_i
AUDIO
$s_i(t)+n(t)$
101b'
Camera + ADC
101a
101a'
ADC
Video Telephony
Face Location (slow,1fps)
Lip Movement (fast,15fps)
Speaker Detection
FFT
Noise Spectrum Estimation
IFFT
Near-End Talk, Double Talk Detection
102
104a
104
104b
106
106a
106b
106c
106d
106e
108
108a
$\underline{o}_{v,t}$
Recognized Speech Signal $\hat{s}_i(nT)$
Detected Near-End Speaker $\hat{S}_i$
Recognized Speech Signal $\hat{s}_i(nT)$

Fig. 1

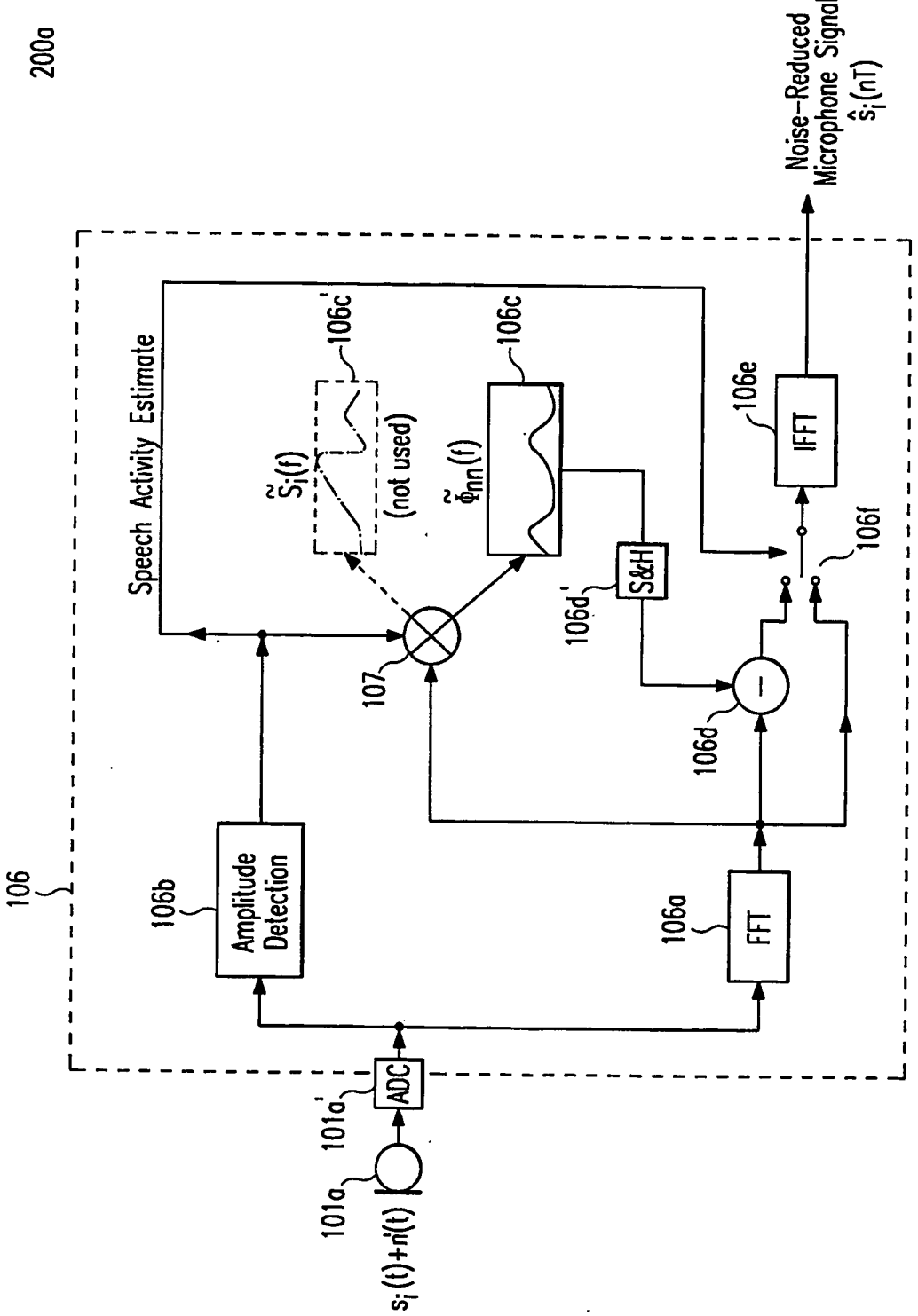

200a
Noise-Reduced Microphone Signal
$\hat{s}_i(nT)$
Speech Activity Estimate
106c'
106c
106e
IFFT
$\tilde{S}_i(f)$
(not used)
$\tilde{\Phi}_{nn}(f)$
106f
S&H
106d'
107
106d
106
106b
Amplitude Detection
106a
FFT
ADC
101a'
101a
$s_i(t)+n_i(t)$

Fig. 2a

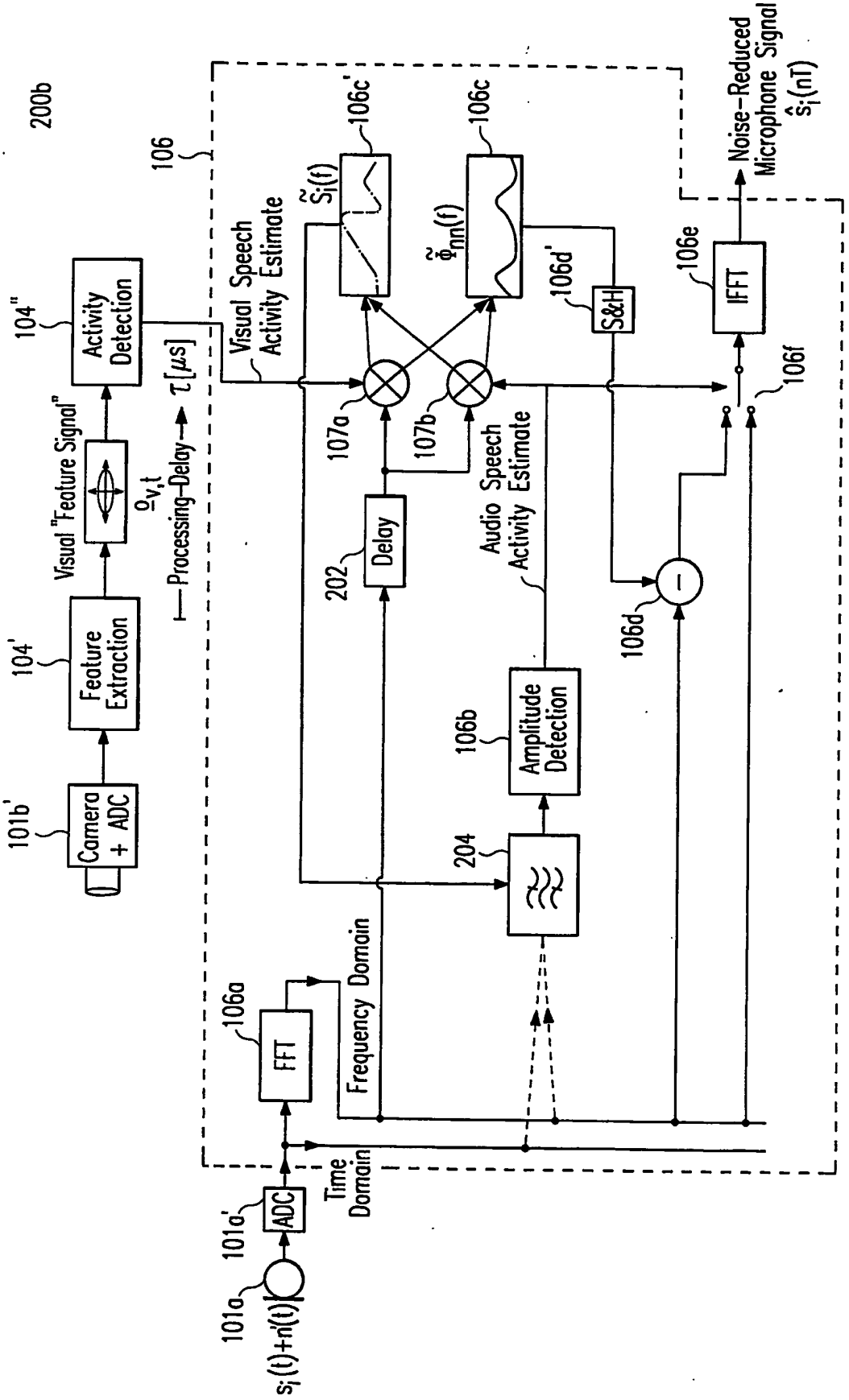

200b
106
Noise-Reduced Microphone Signal
$\hat{s}_i(nT)$
106e
IFFT
106f
106c'
$\tilde{S}_i(f)$
106c
$\tilde{\Phi}_{nn}(f)$
106d'
S&H
Visual Speech Activity Estimate
104"
Activity Detection
Visual "Feature Signal"
$o_{v,t}$
Processing-Delay
$\tau[\mu s]$
107a
107b
Audio Speech Activity Estimate
202
Delay
106d
104'
Feature Extraction
101b'
Camera + ADC
106b
Amplitude Detection
204
106a
FFT
Frequency Domain
Time Domain
101a'
ADC
101a
$s_i(t)+n'(t)$

Fig. 2b

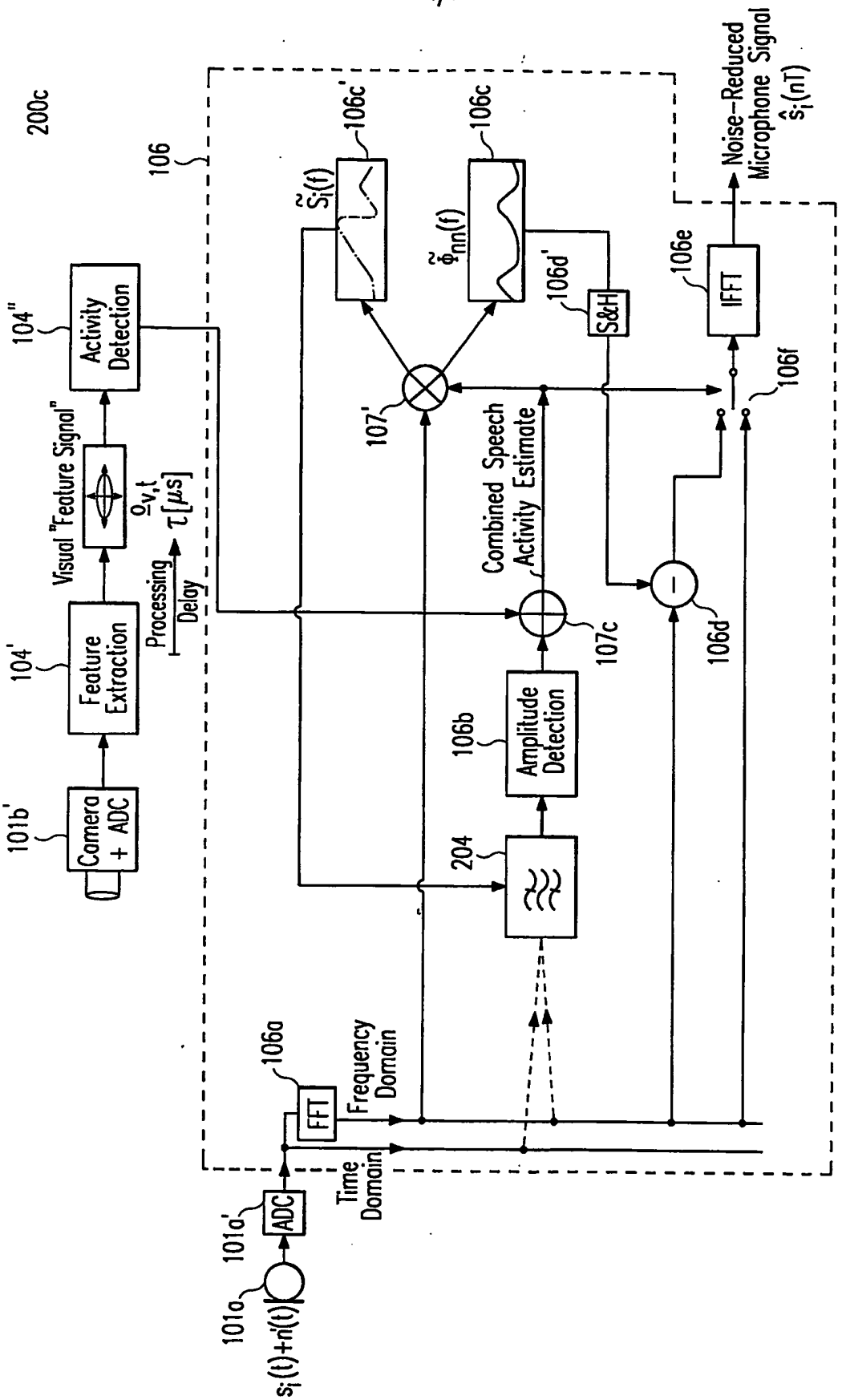

200c
106
101b'
Camera + ADC
104'
Feature Extraction
Visual "Feature Signal"
$\underline{o}_{v,t}$
τ[μs]
Processing Delay
104"
Activity Detection
101a'
ADC
101a
$s_i(t)+n_i(t)$
106a
FFT
Frequency Domain
Time Domain
204
106b
Amplitude Detection
Combined Speech Activity Estimate
107c
107'
$\tilde{S}_i(f)$
106c'
$\tilde{\Phi}_{nn}(f)$
106c
106d'
S&H
106d
106e
IFFT
106f
Noise-Reduced Microphone Signal
$\hat{s}_i(nT)$

Fig. 2c

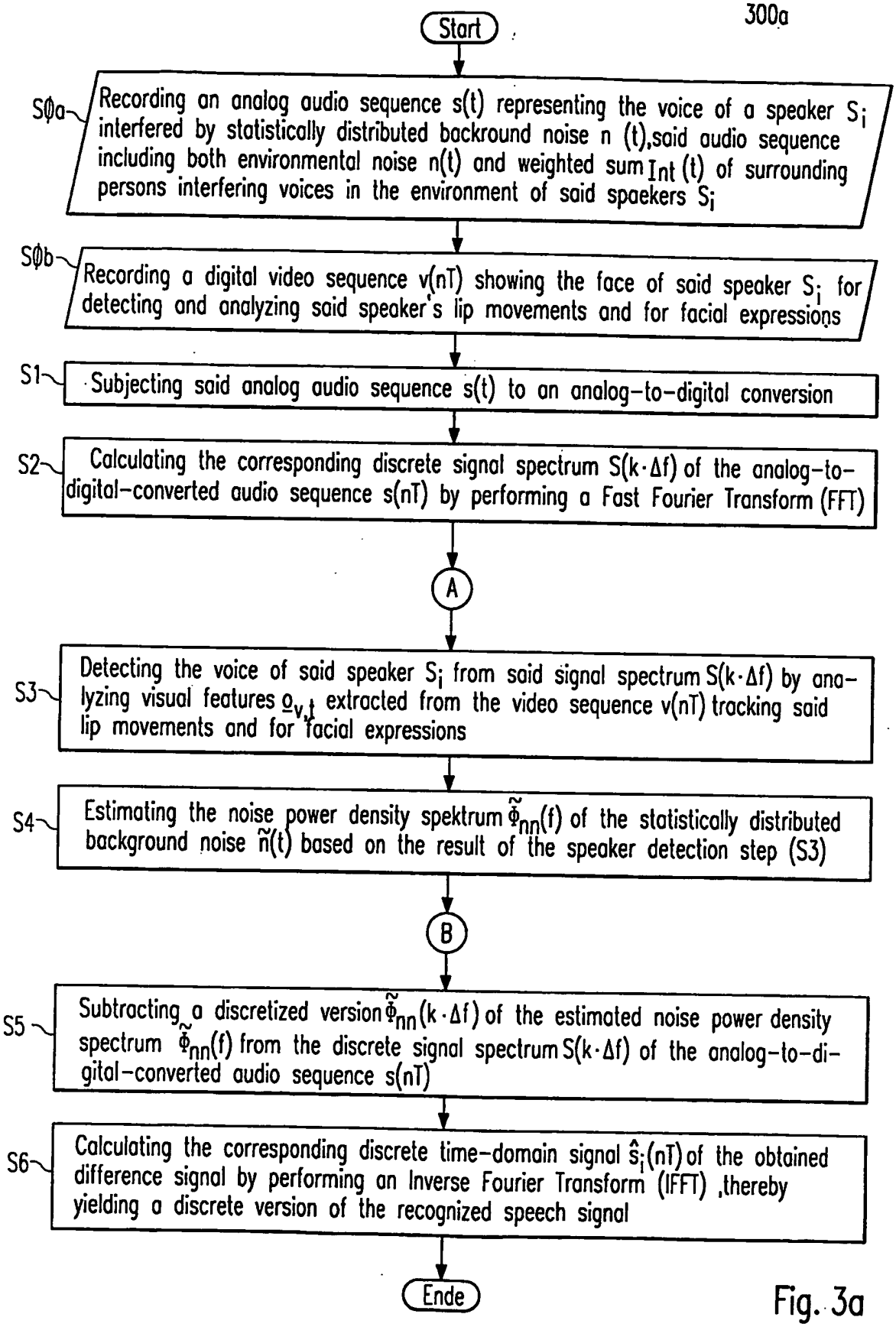

300a
Start
S0a
Recording an analog audio sequence s(t) representing the voice of a speaker S_i interfered by statistically distributed backround noise n (t),said audio sequence including both environmental noise n(t) and weighted sum I_{nt} (t) of surrounding persons interfering voices in the environment of said spaekers S_i
S0b
Recording a digital video sequence v(nT) showing the face of said speaker S_i for detecting and analyzing said speaker's lip movements and for facial expressions
S1
Subjecting said analog audio sequence s(t) to an analog-to-digital conversion
S2
Calculating the corresponding discrete signal spectrum $S(k \cdot \Delta f)$ of the analog-to-digital-converted audio sequence s(nT) by performing a Fast Fourier Transform (FFT)
A
S3
Detecting the voice of said speaker S_i from said signal spectrum $S(k \cdot \Delta f)$ by analyzing visual features $o_{v,t}$ extracted from the video sequence v(nT) tracking said lip movements and for facial expressions
S4
Estimating the noise power density spektrum $\tilde{\Phi}_{nn}(f)$ of the statistically distributed background noise $\tilde{n}(t)$ based on the result of the speaker detection step (S3)
B
S5
Subtracting a discretized version $\tilde{\Phi}_{nn}(k \cdot \Delta f)$ of the estimated noise power density spectrum $\tilde{\Phi}_{nn}(f)$ from the discrete signal spectrum $S(k \cdot \Delta f)$ of the analog-to-digital-converted audio sequence s(nT)
S6
Calculating the corresponding discrete time-domain signal $\hat{s}_i(nT)$ of the obtained difference signal by performing an Inverse Fourier Transform (IFFT) ,thereby yielding a discrete version of the recognized speech signal
Ende

Fig. 3a

300b

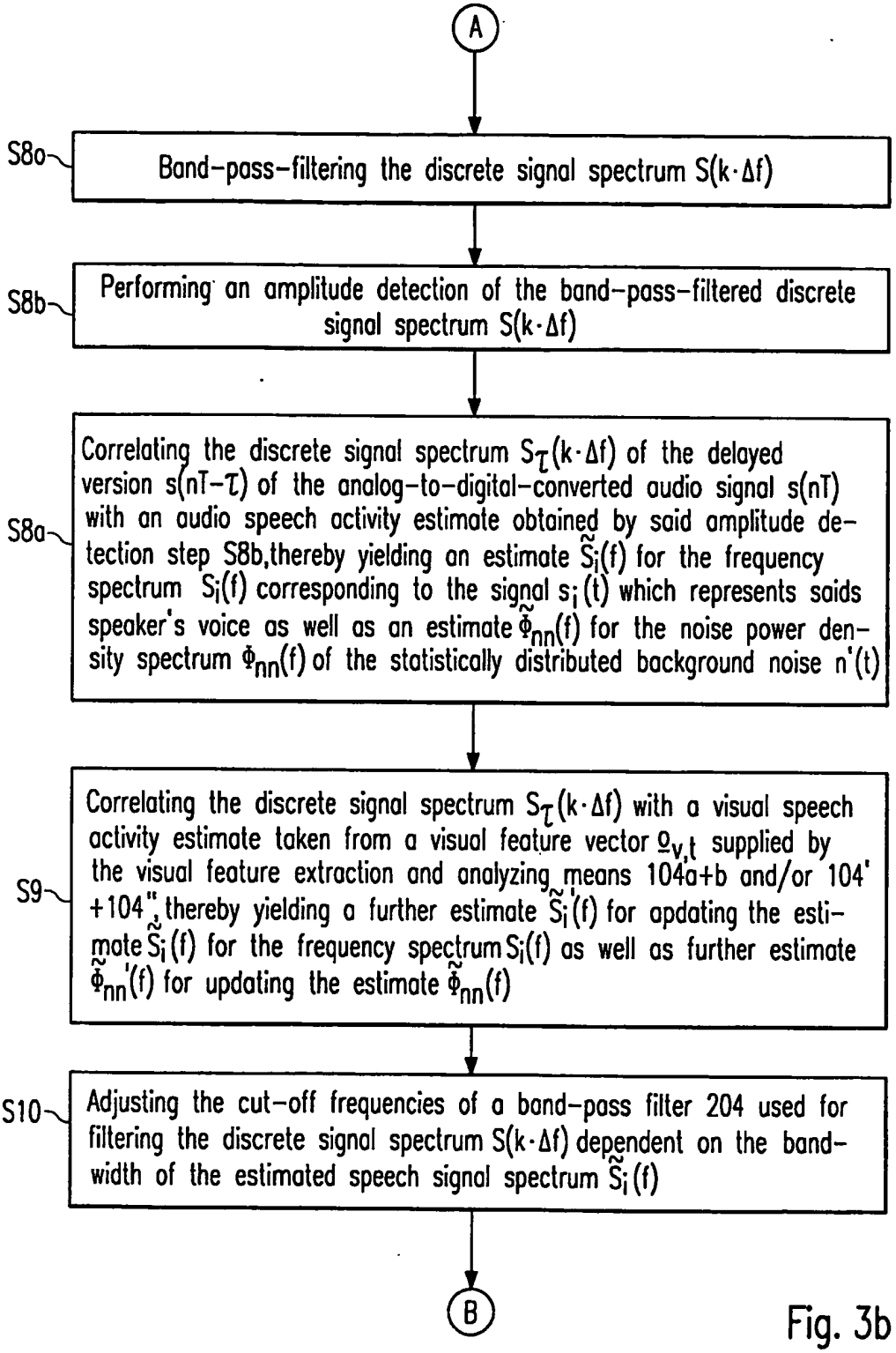

A
S8o
Band-pass-filtering the discrete signal spectrum S(k·Δf)
S8b
Performing an amplitude detection of the band-pass-filtered discrete signal spectrum S(k·Δf)
S8a
Correlating the discrete signal spectrum $S_\tau(k\cdot\Delta f)$ of the delayed version $s(nT-\tau)$ of the analog-to-digital-converted audio signal s(nT) with an audio speech activity estimate obtained by said amplitude detection step S8b, thereby yielding an estimate $\tilde{S}_i(f)$ for the frequency spectrum $S_i(f)$ corresponding to the signal $s_i(t)$ which represents saids speaker's voice as well as an estimate $\tilde{\Phi}_{nn}(f)$ for the noise power density spectrum $\Phi_{nn}(f)$ of the statistically distributed background noise n'(t)
S9
Correlating the discrete signal spectrum $S_\tau(k\cdot\Delta f)$ with a visual speech activity estimate taken from a visual feature vector $\varrho_{v,t}$ supplied by the visual feature extraction and analyzing means 104a+b and/or 104'+104'', thereby yielding a further estimate $\tilde{S}_i'(f)$ for updating the estimate $\tilde{S}_i(f)$ for the frequency spectrum $S_i(f)$ as well as further estimate $\tilde{\Phi}_{nn}'(f)$ for updating the estimate $\tilde{\Phi}_{nn}(f)$
S10
Adjusting the cut-off frequencies of a band-pass filter 204 used for filtering the discrete signal spectrum S(k·Δf) dependent on the bandwidth of the estimated speech signal spectrum $\tilde{S}_i(f)$
B

Fig. 3b

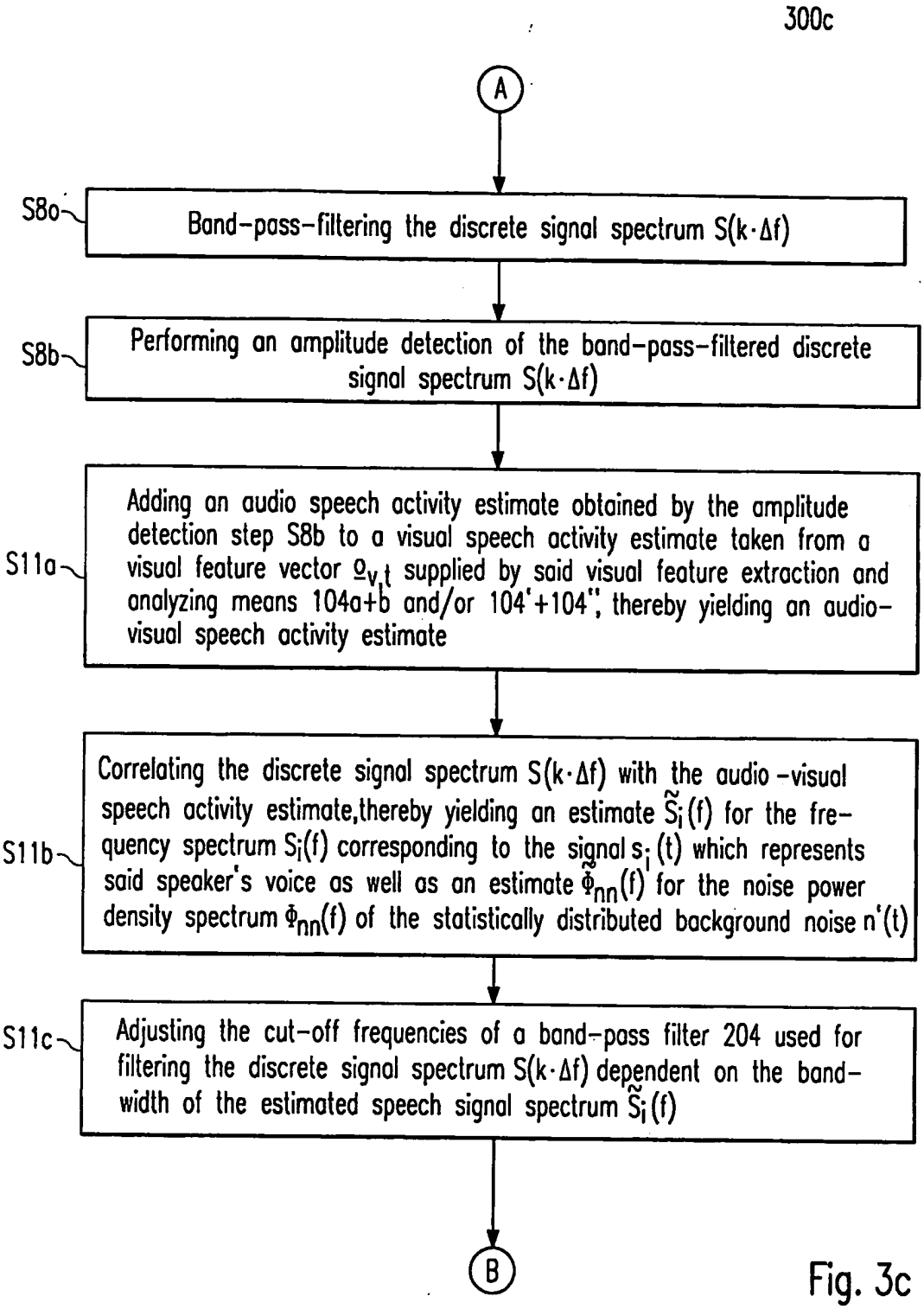

300c
A
S8o
Band-pass-filtering the discrete signal spectrum S(k·Δf)
S8b
Performing an amplitude detection of the band-pass-filtered discrete signal spectrum S(k·Δf)
S11a
Adding an audio speech activity estimate obtained by the amplitude detection step S8b to a visual speech activity estimate taken from a visual feature vector $\underline{o}_{v,t}$ supplied by said visual feature extraction and analyzing means 104a+b and/or 104'+104", thereby yielding an audio-visual speech activity estimate
S11b
Correlating the discrete signal spectrum S(k·Δf) with the audio-visual speech activity estimate,thereby yielding an estimate $\tilde{S}_i(f)$ for the frequency spectrum $S_i(f)$ corresponding to the signal $s_i(t)$ which represents said speaker's voice as well as an estimate $\tilde{\Phi}_{nn}(f)$ for the noise power density spectrum $\Phi_{nn}(f)$ of the statistically distributed background noise n'(t)
S11c
Adjusting the cut-off frequencies of a band-pass filter 204 used for filtering the discrete signal spectrum S(k·Δf) dependent on the bandwidth of the estimated speech signal spectrum $\tilde{S}_i(f)$
B

Fig. 3c

INTERNATIONAL SEARCH REPORT

International Application No
PCT/EP2004/000104

A. CLASSIFICATION OF SUBJECT MATTER
IPC 7 G10L21/02 G10L11/02 G10L15/24

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 7 G10L

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)
EPO-Internal, WPI Data, INSPEC

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X Y	WO 02/29784 A (CLARITY LLC ; ERTEN GAMZE (US)) 11 April 2002 (2002-04-11) abstract; figures 11,13-15	1,2,4, 14,15 3,5-9
Y	WO 02/084644 A (DEUTSCHE TELECOM AG) 24 October 2002 (2002-10-24) abstract & US 2003/191633 A1 (BERGER JENS) 9 October 2003 (2003-10-09) abstract	5,8,9
Y	US 2003/007633 A1 (WILDIE MARK GREIG ET AL) 9 January 2003 (2003-01-09) abstract	3,6,7

-/--

[X] Further documents are listed in the continuation of box C.

[X] Patent family members are listed in annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
10 May 2004	28/05/2004

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Quélavoine, R

Form PCT/ISA/210 (second sheet) (January 2004)

INTERNATIONAL SEARCH REPORT

International Application No
PCT/EP2004/000104

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	GIRIN L ET AL: "NOISY SPEECH ENHANCEMENT BY FUSION OF AUDITORY AND VISUAL INFORMATION: A STUDY OF VOWEL TRANSITIONS" 5TH EUROPEAN CONFERENCE ON SPEECH COMMUNICATION AND TECHNOLOGY. EUROSPEECH '97. RHODES, GREECE, SEPT. 22 - 25, 1997, EUROPEAN CONFERENCE ON SPEECH COMMUNICATION AND TECHNOLOGY. (EUROSPEECH), GRENOBLE : ESCA, FR, vol. VOL. 5 OF 5, 22 September 1997 (1997-09-22), pages 2555-2558, XP001045210 the whole document	1,2,4

INTERNATIONAL SEARCH REPORT

Information on patent family members

International Application No
PCT/EP2004/000104

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 0229784 A	11-04-2002	AU 9645901 A	15-04-2002
		WO 0229784 A1	11-04-2002
		US 2002116197 A1	22-08-2002
WO 02084644 A	24-10-2002	DE 10120168 A1	24-10-2002
		WO 02084644 A1	24-10-2002
		EP 1382034 A1	21-01-2004
		US 2003191633 A1	09-10-2003
US 2003191633 A1	09-10-2003	DE 10120168 A1	24-10-2002
		WO 02084644 A1	24-10-2002
		EP 1382034 A1	21-01-2004
US 2003007633 A1	09-01-2003	NONE	